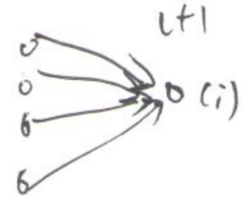
fp: $x \rightarrow$ 

a_j^l z_i^{l+1}

$j=1\cdots s_l$ $i=1,\cdots s_{l+1}$

对于第l+1层的第i个neuron，其来源

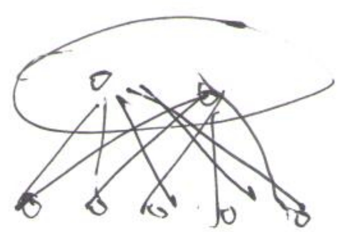
$$z_i^{l+1} = \sum_{j=1}^{s_l} W_{ij} \cdot a_j^l$$



$$a_i^{l+1} = f(z_i^{l+1})$$

到了n_L层（即output layer）

以2类分类问题为例：



一样的做法.

现对n_L层用softmax.（这时一般用cross entropy loss）

或者也不用softmax，直接square-error.

① BP $J(w,b) = \left[\frac{1}{m}\sum_{i=1}^{m} J(w,b;x^{(i)},y^{(i)})\right] + \frac{\lambda}{2}\underbrace{\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(w_{ji}^{(l)})^2}$

m-traing set. $\{(x^{(1)},y^{(1)}), \dots (x^{(m)},y^{(m)})\}$

$$= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\|h_{w,b}(x^{(i)}) - y^{(i)}\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(w_{ji}^{(l)})^2$$

②

$$W_{ij}^{(l)} = w_{ij}^{(l)} - \alpha\frac{\partial}{\partial w_{ij}^{(l)}}J(w,b)$$

$$b_i^{(l)} = b_i^{(l)} - \alpha\frac{\partial}{\partial b_i^{(l)}}J(w,b)$$

③

$$\frac{\partial J(w,b)}{\partial W_{ij}^{(l)}} = \frac{1}{m}\left[\sum_{i=1}^{m}\frac{\partial}{\partial w_{ij}^{(l)}}J(w,b;x,y)\right] + \lambda w_{ij}^{(l)}$$

$$\frac{\partial}{\partial b_i^{l}}J(w,b) = \left[\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial b_i^{l}}J(w,b;x,y)\right]$$

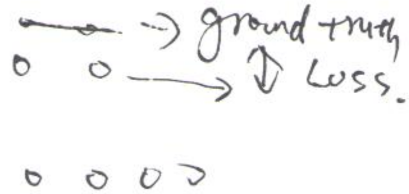
④ 关键就是求 $\frac{\partial}{\partial w_{ij}^{(l)}}J(w,b;x,y)$ 和 $\frac{\partial}{\partial b_i^{l}}J(w,b;x,y)$

4.1 for the output layer, $loss = \frac{1}{2}\|y - h_{w,b}(x)\|^2$

$h_{w,b}(x) = a_i^{n_l} = f(z_i^{n_l})$

$$\frac{\partial \frac{1}{2}\|y_i - h_{w,b}(x)\|^2}{\partial z_i^{(n_l)}} = \frac{\partial \frac{1}{2}\|y_i - f(z_i^{n_l})\|^2}{n_l} = (y_i - f(z_i^{n_l}))(-f'(z_i^{n_l}))$$

BP过程.



$$J(W,b) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}\|h_{W,b}(x^{(i)}) - y^{(i)}\|^2$$

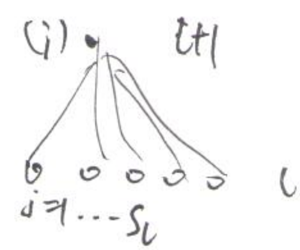
$$+ \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}} \left(W_{ji}^{(l)}\right)^2 \to \text{weight decay part.}$$

In example.

4.2. for $i = n_l-1, n_l-2, \ldots, 2$.

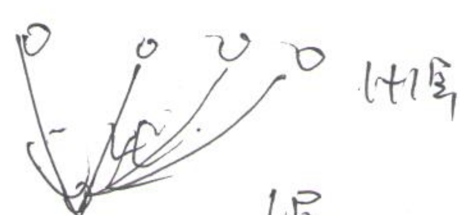
$$z_i^{l+1} = \sum_{j=1}^{s_l} W_{ij}^{l} a_j^{l}$$

$$a_i^{l+1} = f(z_i^{l+1})$$



现在定义一个概念. sensitivity. $\delta_i^{(l)}$ 表达的是第l层的第i个neuron 的activation. 对output的loss的responsibility. 如果这个 responsibility 很大, 说明 a_i^l 对于的loss很重要, vice versa, 所以 $\delta_i^{(l)}$ 为 sensitivity neuron 的 activation 对于 general loss 的 sensitivity.

$$\delta_i^{(l)} = \sum_{j=1}^{s_{l+1}} \left(W_{ij}^{(l)} \delta_j^{(l+1)}\right) f'(z_i^{l})$$



这样便定义了"error" of l's layer to l+1's layer.

4.3. $\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x,y) = a_j^{(l)} \cdot \delta_i^{(l+1)}$

$\frac{\partial}{\partial b_i^{(l)}} J(W,b;x,y) = \delta_i^{(l+1)}$

$$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial a}\frac{\partial a}{\partial W} = \delta_i^{(l+1)} \cdot a_j^{(l)}$$

$$\frac{\partial J}{\partial b} = \frac{\partial J}{\partial a}\cdot\frac{\partial a}{\partial b}$$